

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開平7-160436

(43) 公開日 平成7年(1995)6月23日

(51) Int.Cl. <sup>6</sup>	識別記号	庁内整理番号	F I	技術表示箇所
G 0 6 F 3/06	5 4 0	3 0 2 A		

審査請求 未請求 請求項の数 8 O L (全 7 頁)

(21) 出願番号 特願平5-310612

(22) 出願日 平成5年(1993)12月10日

(71) 出願人 000005108

株式会社日立製作所

東京都千代田区神田駿河台四丁目6番地

(72) 発明者 田中 幸一

神奈川県小田原市国府津2880番地 株式会

社日立製作所ストレージシステム事業部内

(72) 発明者 村岡 健司

神奈川県小田原市国府津2880番地 株式会

社日立製作所ストレージシステム事業部内

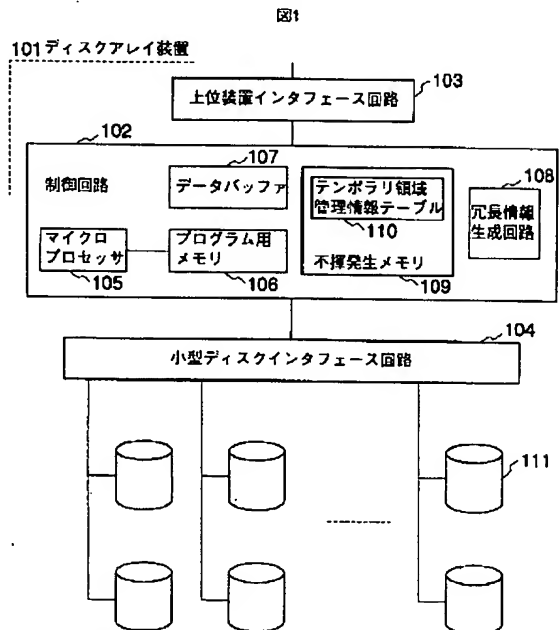
(74) 代理人 弁理士 富田 和子

(54) 【発明の名称】 ディスクアレイ装置

(57) 【要約】

【目的】 コストを上昇させることなく、書き込み処理時の性能低下（ライトペナルティ）を防いだディスクアレイ装置を提供する。

【構成】 ディスク装置111上の格納データを書き込む領域が、格納データを一時的に書き込むためのテンポラリ領域と、格納データを最終的に書き込むための格納データ領域とに分割されている。マイクロプロセッサ105は、外部からの格納データは、冗長データを作成することなく、一旦、テンポラリ領域に書き込み、テンポラリ領域への書き込みが完了した時点で、外部へ書き込み完了を報告する。書き込み完了の報告後、テンポラリ領域上の格納データに基づいて、冗長情報生成回路110は、冗長データを生成し、マイクロプロセッサ105は、冗長データを冗長データ領域に格納し、格納データを格納データ領域に書き込む。



111: 磁気ディスク装置

## 【特許請求の範囲】

【請求項 1】複数のディスク装置を備え、当該複数のディスク装置のうち少なくとも 1 つは、他のディスク装置に書き込まれた格納データを読み出すことができないときに、読み出すことができない当該データを再現するための冗長データを書き込む冗長データ領域を有するディスクアレイ装置において、

上記ディスク装置のうちの少なくとも 1 つは、上記格納すべき格納データを書き込む領域が、当該格納データを一時的に書き込むためのテンポラリ領域と、当該テンポラリ領域に書き込まれた当該格納データを最終的に書き込むための格納データ領域とに分割されており、

外部から受付けた、上記ディスク装置に格納すべき格納データは、冗長データを作成することなく、一旦、上記テンポラリ領域に書き込み、当該テンポラリ領域への書き込みが完了した時点で、外部へ書き込み完了を報告するテンポラリ領域書き込み手段と、

上記書き込み完了の報告後、テンポラリ領域上の上記格納データに基づいて、冗長データを生成する生成手段と、

当該冗長データを上記冗長データ領域に格納し、当該格納データを上記格納データ領域に書き込むデータ領域書き込み手段とを有することを特徴とするディスクアレイ装置。

【請求項 2】請求項 1 記載のディスクアレイ装置において、

外部から格納すべき格納データを受付けたときに、当該格納データを書き込むテンポラリ領域を、最終的に当該格納データを書き込むべき格納データ領域を有するディスク装置とは異なるディスク装置にあるテンポラリ領域から選択する手段を有することを特徴とするディスクアレイ装置。

【請求項 3】請求項 1 記載のディスクアレイ装置において、

外部から格納すべき格納データを受付けたときに、当該格納データを書き込むテンポラリ領域を、読み出しもしくは書き込み動作を行っていないディスク装置にあるテンポラリ領域から選択する手段を有することを特徴とするディスクアレイ装置。

【請求項 4】請求項 1 記載のディスクアレイ装置において、

外部から格納すべき格納データを受付けたときに、複数の上記ディスク装置を選択し、当該格納データを書き込むテンポラリ領域を、当該ディスク装置にあるテンポラリ領域とし、複数のテンポラリ領域の各々に同一の格納データを多重に書き込む手段と、

上記テンポラリ領域から上記格納データを読み出す際に、一方のディスク装置から上記格納データを読み出すことができないときは、他方のディスク装置から読み出す手段とを有することを特徴とするディスクアレイ装

置。

【請求項 5】請求項 4 記載のディスクアレイ装置において、

一方のディスク装置から上記格納データを読み出すことができないときは、当該ディスク装置が故障しているときであることを特徴とするディスクアレイ装置。

【請求項 6】請求項 4 記載のディスクアレイ装置において、

一方のディスク装置から上記格納データを読み出すことができないときは、当該ディスク装置が、読み出しもしくは書き込み動作を行っているときであることを特徴とするディスクアレイ装置。

【請求項 7】請求項 1 記載のディスクアレイ装置において、

あらかじめ定められたデータ長よりも、上記格納データのデータ長が大きい場合には、上記テンポラリ領域に格納しないで、当該格納データに基づいて、冗長データを生成する生成手段と、

当該冗長データを上記冗長データ領域に格納し、当該格納データを上記格納データ領域に書き込む手段とを有することを特徴とするディスクアレイ装置。

【請求項 8】請求項 1 記載のディスクアレイ装置において、

上記テンポラリ領域から上記格納データを読みだし、上記格納データ領域へ格納データを書き込む際に、上記格納データ領域を有するディスク装置で書き込みエラーが発生したときは、当該ディスク装置を切り離し、当該ディスク装置への書き込みは以後行わないことを特徴とするディスクアレイ装置。

【発明の詳細な説明】

【0001】

【産業上の利用分野】本発明は、いわゆる RAID (Redundant Array of Inexpensive Disks) と呼ばれるディスクアレイ装置に係わり、ディスク装置のアクセス処理性能、特に、データ書き込み処理の高速化に関する。

【0002】

【従来の技術】コンピュータを構成する要素のうち、駆動機構を有するハードディスク装置（ディスク装置またはドライブとも呼ばれる）は、他の電子部品に比べ故障率が高く、また、データを書き込む手段として用いられるため、その故障が与える影響は、大きい。そのため、信頼性の向上が要望されていた。そのような要望に応えるものとして、特開平 2-236714 号公報に記載されているものがある。これは、複数のディスク装置を有し、書き込むべき格納データの書き込み時には、格納データの書き込みに加えて、冗長データを生成し、書き込む。冗長データは、1 台のディスク装置が故障して、そのディスク装置上の格納データが読めなくなったとき

に、他の故障していないディスク装置上の格納データと

## 3

冗長データとから読めなくなったデータが再現できるように生成される。生成された冗長データは、当該冗長データを生成するのに使われた格納データを格納しているディスク装置とは異なる、別のディスク装置へ書き込まれる。このように構成された、複数のディスク装置からなるものは、ディスクアレイ装置、もしくは、RAIDと呼ばれる。これによれば、アレイを構成する任意のディスク装置が故障しても、他のディスク装置の格納データと前記冗長データから該障害ディスク装置上の格納データを修復でき、ハードディスク装置のデータ保全性の向上が図れる。

## 【0003】

【発明が解決しようとする課題】RAIDは、データ保全性が向上する半面、格納データの書き込みには、常により冗長データの生成と書き込みとを伴い、そのために書き込み処理時間が増大し、書き込み完了の報告を外部にするのが遅れて、性能が低下してしまうという問題点が指摘されていた。そのため、ディスク装置への書き込みが少ない（読み込み処理が多くの割合を占める）システムにおいて、事実上向かないものであった。この対策として、ディスク制御装置内にキャッシュメモリを用意し、キャッシュメモリへの書き込みが完了すると、ただちにホストコンピュータに書き込み完了を報告し、その後、冗長データを生成し、格納データと冗長データをディスク装置に書き込むシステムもある。しかし、このシステムは、高価なキャッシュメモリを用いるために、システムが高価になるという問題がある。本発明の目的は、コストを上昇させることなく、上記書き込み処理時の性能低下（ライトペナルティ）を防いだディスクアレイ装置を提供することにある。本発明の他の目的は、上記の目的に加えて、RAIDの特徴であるデータの保全性も維持できるディスクアレイ装置を提供することにある。

## 【0004】

【課題を解決するための手段】上記目的を達成するために、複数のディスク装置を備え、当該複数のディスク装置のうち少なくとも1つは、他のディスク装置に書き込まれた格納データを読み出すことができないときに、読み出すことができない当該データを再現するための冗長データを書き込む冗長データ領域を有するディスクアレイ装置において、上記ディスク装置のうちの少なくとも1つは、上記格納すべき格納データを書き込む領域が、当該格納データを一時的に書き込むためのテンポラリ領域と、当該テンポラリ領域に書き込まれた当該格納データを最終的に書き込むための格納データ領域とに分割されており、外部から受付けた、上記ディスク装置に書き込むべき格納データは、冗長データを作成することなく、一旦、上記テンポラリ領域に書き込み、当該テンポラリ領域への書き込みが完了した時点で、外部へ書き込み完了を報告するテンポラリ領域書き込み手段と、上記

## 4

書き込み完了の報告後、テンポラリ領域上の上記格納データに基づいて、冗長データを生成する生成手段と、当該冗長データを上記冗長データ領域に格納し、当該格納データを上記格納データ領域に書き込むデータ領域書き込み手段とを有することとしたものである。また、外部から格納すべき格納データを受付けたときに、複数の上記ディスク装置を選択し、当該格納データを書き込むテンポラリ領域を、当該ディスク装置にあるテンポラリ領域とし、複数のテンポラリ領域の各々に同一の格納データを多重に書き込む手段と、上記テンポラリ領域から上記格納データを読み出す際に、一方のディスク装置から上記格納データを読み出すことができないときは、他方のディスク装置から読み出す手段とを有することとしたものである。また、外部から格納すべき格納データを受付けたときに、当該格納データを書き込むテンポラリ領域を、最終的に当該格納データを書き込むべき格納データ領域を有するディスク装置とは異なるディスク装置にあるテンポラリ領域から選択する手段を有することとしたものである。

## 【0005】

【作用】外部からの格納すべき格納データは、冗長データを作成することなく一旦テンポラリ領域に書き込み、この書き込みが完了した時点で外部（例えば、ホストコンピュータ）へ書き込み完了を報告し、テンポラリ領域上の格納データを、その後読み出して、冗長データを生成し、格納データと、冗長データとをそれぞれのデータ領域に書き込む。これにより、キャッシュメモリを用いずに、外部へのレスポンスが高速化される。また、複数のディスク装置上のテンポラリ領域に2重書きをし、その後、同一格納データへの読み込み要求を受けた場合、テンポラリ領域から該当格納データを読み込むこととすると、格納データの保全性が高まる。このときに、テンポラリ領域からの読み出しにおいて、2重書きしているディスク装置のうちで、現在読み出し／書き込み動作を行っていないディスク装置（このディスク装置はドライブビジーになっていないと呼ばれる）を選択することにより、ドライブビジーによる遅延を回避し、高速に書き込み動作を行うことができる。また、テンポラリ領域から格納データ領域に格納データを書き込む際に、古い格納データとテンポラリ領域にある新しい格納データと冗長データとを読みだして、新たに冗長データを生成する必要がある。当該格納データを書き込むテンポラリ領域を、最終的に当該格納データを書き込むべき格納データ領域を有するディスク装置とは異なるディスク装置にあるテンポラリ領域から選択すると、各ドライブを多重動作できることにより、テンポラリ領域から格納データ領域への書き込みが高速化する。このときに、選択するテンポラリ領域として、現在読み出し／書き込み動作を行っていないディスク装置にあるものから選択することにより、ドライブビジーによる遅延を回避し、高速に書き

込み動作を行える。

【0006】

【実施例】以下、本発明に係わるディスクアレイ装置の一実施例について説明する。図1は、本実施例に係わるディスクアレイ装置の全体構成を示した図である。図中ディスクアレイ装置101は、制御回路102、上位インタフェース回路103、小型ディスクインタフェース回路104、磁気ディスク装置111より構成される。また制御回路102は、マイクロプロセッサ105、プログラム用メモリ106、データバッファ107、冗長情報生成回路108、不揮発性メモリ109から構成されている。また、不揮発性メモリ109は、テンポラリ領域管理テーブル110を有する。図2は、ディスクアレイ装置内の磁気ディスク装置の構成を示した図である。ディスクアレイ装置101内の全ての磁気ディスク装置111は、ユーザデータを最終的に格納するデータ領域201と、ユーザデータを一時的に書き込むテンポラリ領域202より構成される。データ領域201とテンポラリ領域202のサイズは任意である。図3は、本実施例に係わるソフトウェア構成図である。これは、図1に示す、マイクロプロセッサ105と、マイクロプロセッサ105が実行するマイクロプログラムが格納されているプログラム用メモリ106とにより実行される。ただし、バックグラウンドデータ書き込み処理304のうち冗長データの生成は、冗長情報生成回路108が行う。スケジューラ301によりディスクアレイ装置101全体が制御される。スケジューラ301によりホストI/F制御部302、ホストコマンド実行部303、バックグラウンドデータ書き込み処理部304は制御される。また、ホストコマンド実行部303は、コマンド判定部305を通して、リードコマンド処理306、ライトコマンド処理307、制御系コマンド処理308を利用することによりホストコマンドを実行する。図4は、本実施例に係わるテンポラリ領域管理テーブル110の構成を示した図である。図中、本管理テーブルは、上位装置指定情報401、制御部管理情報402、テンポラリ領域管理情報403、およびテンポラリ領域のうち使われているものと空いているものを確認するためのネクストポインタ404から構成されている。ネクストポインタ404には、2種類のチェーンがあり、1つは、テンポラリ領域のうち使われているものを結んだチェーンであり、もう1つは、空いているものを結んだチェーンである。上位装置指定情報401は、論理LBA(LBA: Logical Block Address)405、論理LEN(LEN: Length)406から構成される。制御部管理情報402は、物理ドライブNo. 407、物理LBA408、物理LEN409から構成される。テンポラリ領域管理情報403は、テンポラリ領域として使用される物理ドライブNo. 410、その磁気ディスク装置内の物理LBA411、および物

理LEN412から構成される。ネクストポインタ404を確認することによりテンポラリ領域の空きを確認することが可能である。次に上位装置からライトコマンドを受取り、テンポラリ領域に対して2重書きを行う各処理について図5、図7を用いて説明する。上位装置501からライトデータ502を受け取ると、制御部503は、まずライトデータ502のブロック長を確認し(図7、701)、しきい値よりブロック長が大きかった場合は、テンポラリ領域に格納せずに、通常のライト処理(702)を実施する。つまり上位装置501からのライトデータ502は、通常のライト領域508へ格納し、また、作成されたパリティデータは、通常のパリティライト領域509へ書き込む。ライトデータ502のサイズが大きいときは、ライトパナルティは小さいから直接、最終的な格納場所に格納することとしたものである。しきい値よりブロック長が小さい場合は、テンポラリ領域をネクストポインタ404(図4)を使用して確保する(703)。確保する際には、実際にデータを書く物理的な磁気ディスク装置と重ならないように管理している。異なるディスク装置とすることにより、バックグラウンドで処理するとき、テンポラリ領域にある新データと実際にデータを書く位置にある旧データとが異なるディスク装置であるために、並列に読みだすことができ、冗長データを生成する処理が高速化する。さらに、テンポラリ領域を確保する際は、格納データを書き込むテンポラリ領域を、現に読み出しもしくは書き込み動作を行っていないディスク装置にあるテンポラリ領域から選択するようにする。こうすることにより、ドライビジーによる遅延を回避し、高速に書き込み動作を行うことができる。もし、テンポラリ領域に空きが無く確保できなかった場合(704)には、通常のライト処理(702)を実施する。空きテンポラリ領域が確保できた場合、テンポラリ領域511のうちの領域506/507に対して2重書きを実施する。次にテンポラリ領域に対して2重書きしているユーザデータを読み込み、パリティデータを生成し、データ領域に書き込む処理について図6、図8を用いて説明する。制御部603は、バックグラウンド処理で、まずテンポラリ領域管理テーブル611よりすでに2重書きされているユーザデータが格納されている領域605/606のアドレスを確保し(図8、801)、今現在リードおよびライト処理を行っていない方のテンポラリ領域から、データをリードする(802)。つまり、2重書きされているユーザデータが格納されている領域605/606のうち、例えばデータが格納されている領域606が格納されている磁気ディスク装置がリード処理で使用されている場合、別のデータが格納されている領域605をリードすることになる。こうすることにより、ドライビジーによる遅延を回避し、高速に読み出し動作を行うことができる。磁気ディスク装置の障害により1つのテンポラリ領域か

らデータがリードできなかった場合(803)には、2重書きしている別のテンポラリ領域からデータをリードする(804)。このように、2重書きしていることにより、ディスク装置が故障してもデータを保持することができ、データの保全性が高まる。テンポラリ領域から読み出されたデータと、変更する必要があるパリティデータ608により、冗長情報生成回路604で新しいパリティデータを生成し(805)、データは、607へ、パリティデータは、608へ書き込まれる(806)。データ領域へ格納データを書き込む場合に、上記格納データ領域を有するディスク装置で書き込みエラーが発生したときは、当該ディスク装置を自動的に切り離し、このディスク装置への書き込みは以後行わないこととする。尚、以上の実施例においては、記憶媒体として磁気ディスク装置を用いたディスクアレイ装置の例を示したが、他の記録媒体、たとえば、光ディスクを用いた場合でも同様に実現できる。

#### 【0007】

【発明の効果】本発明によれば、いわゆるRAID4、5におけるライトペナルティによる書き込み処理性能の低下を防ぐことができるようになり、ディスク装置への書き込みが少ない(読み込み処理が多く割合を占める)システムであるかどうかによらず、RAID利用によるシステムの性能向上を図ることが可能となる。また、従来のOS(UNIX等)と同じインタフェースをとることができ、アプリケーションに対して、テンポラリ領域を意識させないので、既存のアプリケーションを活用することが可能であるという効果がある。

#### 【図面の簡単な説明】

【図1】本発明の一実施例に係わるディスクアレイ装置のブロック図。

【図2】本発明の一実施例に係わる磁気ディスク装置内の格納領域の説明図。

【図3】本発明の一実施例に係わるソフトウェア構成

図。

【図4】本発明の一実施例に係わるテンポラリ領域管理テーブルの説明図。

【図5】本発明の一実施例に係わる書き込み動作の説明図。

【図6】本発明の一実施例に係わるバックグラウンド時の書き込み動作の説明図。

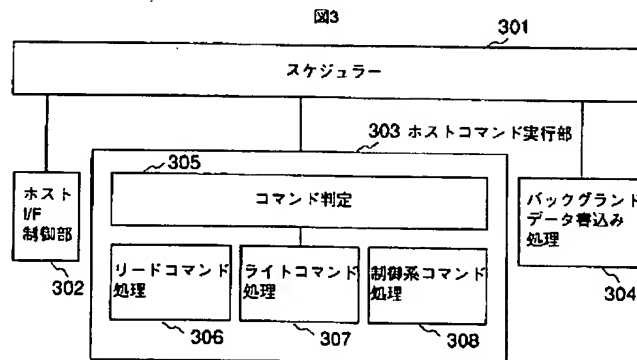
【図7】本発明の一実施例に係わるテンポラリ領域へのデータ書き込み処理のフロー図。

【図8】本発明の一実施例に係わるテンポラリ領域からデータ領域へのデータ書き込み処理のフロー図。

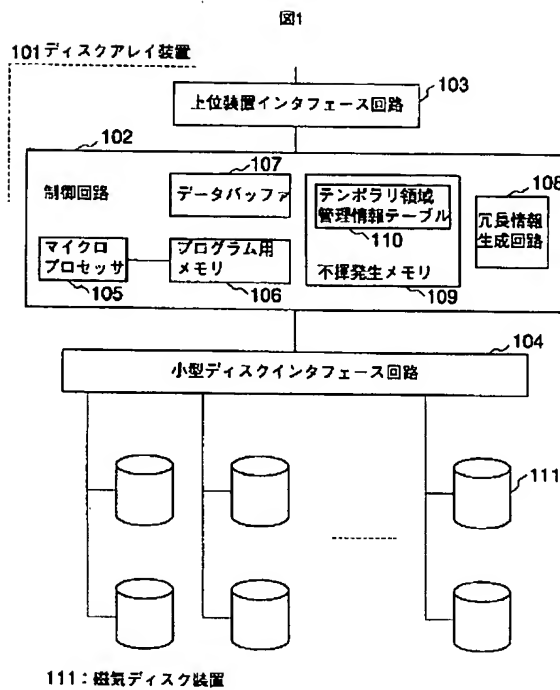
#### 【符号の説明】

- 101 ディスクアレイ装置
- 102 制御回路
- 103 上位インタフェース回路
- 104 小型ディスクインタフェース回路
- 105 マイクロプロセッサ
- 106 プログラム用メモリ
- 107 データバッファ
- 108 冗長情報生成回路
- 109 不揮発性メモリ
- 110 テンポラリ領域管理テーブル
- 111 磁気ディスク装置
- 201 データ領域
- 202 テンポラリ領域
- 301 スケジューラ
- 302 ホストI/F制御部
- 303 ホストコマンド実行部
- 304 バックグラウンドデータ書き込み処理
- 305 コマンド判定
- 306 リードコマンド処理
- 307 ライトコマンド処理
- 308 制御系コマンド処理

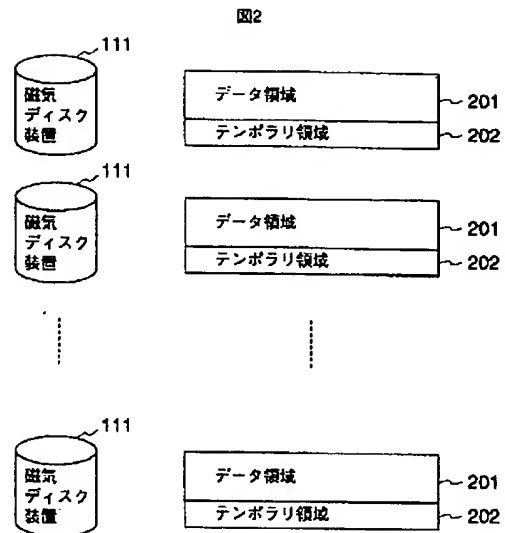
【図3】



【図 1】



【図 2】

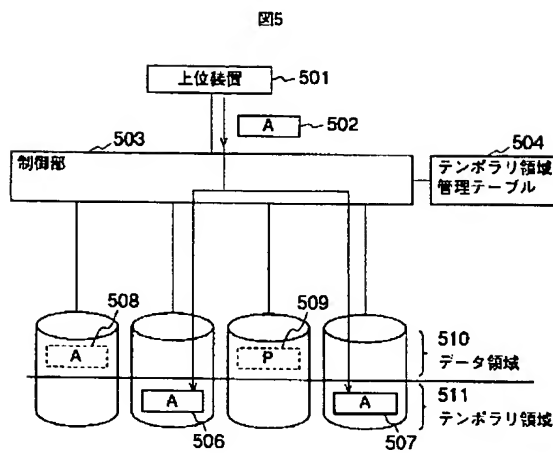


【図 4】

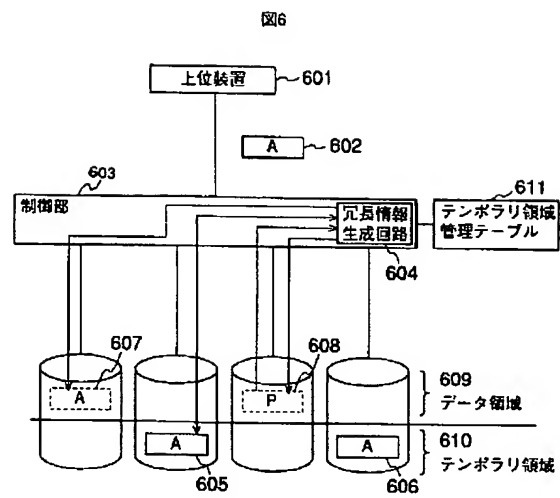
図4

404 ネクスト ポインタ	401 上位装置指定情報		402 制御部管理情報			403 テンポラリ領域情報		
	論理LBA	論理LEN	物理 ドライブ No.	物理LBA	物理LEN	物理 ドライブ No.	物理LBA	物理LEN
	405	406	407	408	409	410	411	412
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

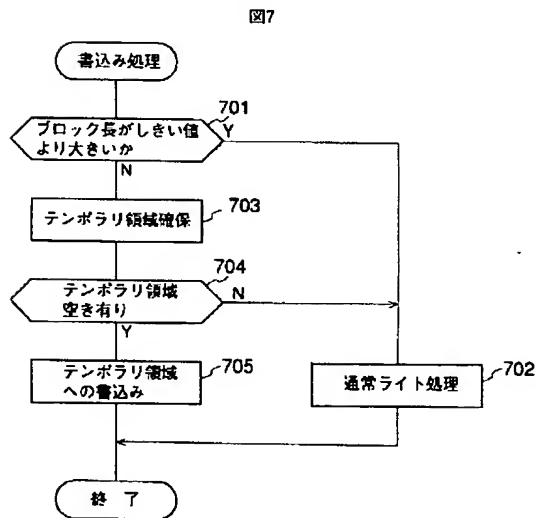
【図5】



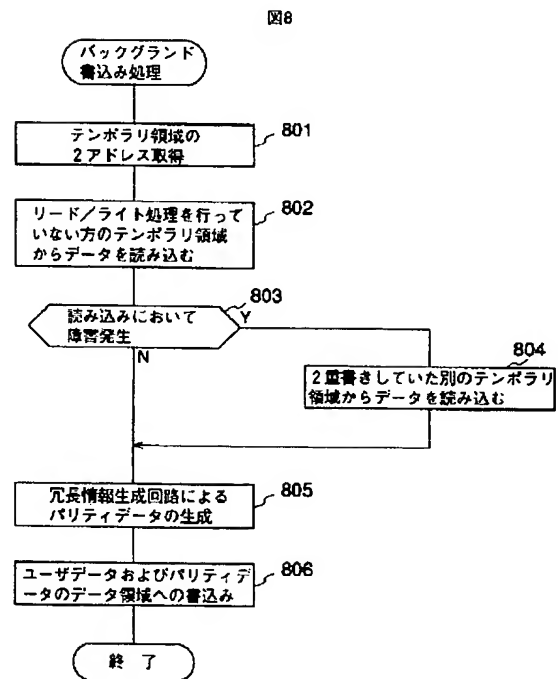
【図6】



【図7】



【図8】



**This Page is Inserted by IFW Indexing and Scanning  
Operations and is not part of the Official Record**

**BEST AVAILABLE IMAGES**

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- ☐ BLACK BORDERS
- ☐ IMAGE CUT OFF AT TOP, BOTTOM OR SIDES
- ☒ FADED TEXT OR DRAWING
- ☐ BLURRED OR ILLEGIBLE TEXT OR DRAWING
- ☐ SKEWED/SLANTED IMAGES
- ☒ COLOR OR BLACK AND WHITE PHOTOGRAPHS
- ☐ GRAY SCALE DOCUMENTS
- ☐ LINES OR MARKS ON ORIGINAL DOCUMENT
- ☐ REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY
- ☐ OTHER: \_\_\_\_\_

**IMAGES ARE BEST AVAILABLE COPY.**

**As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.**